DOCUMENT RESUME

ED 390 932 TM 024 544

AUTHOR Nandakumar, Ratna
TITLE Application of SIBTEST in Dealing with Issues of DIF in the Context of Multidimensional Data.
PUB DATE 19 Apr 95
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Ability; Identification; *Item Bias; Simulation; *Statistical Significance
IDENTIFIERS *Power (Statistics); *SIBTEST (Computer Program)

ABSTRACT

A modification of the SIBTEST procedure to assess differential item functioning (DIF) for two-dimensional test data (i.e., data for tests where two intended abilities are tapped by test items) is described. A small simulation study is carried out to assess the performance of the modified SIBTEST to detect DIF in such two-dimensional data. The simulation study results indicate that the modified SIBTEST has exhibited good adherence to the nominal level of significance for cases where there are two intended abilities with no DIF present, while exhibiting high power for cases where there are two intended abilities with DIF present. The results are discussed with implications for future research. (Contains 2 tables and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 390 932

Application of SIBTEST in Dealing With Issues of DIF in the Context of Multidimensional Data

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

Ratna Nandakumar

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Ratna Nandakumar
Department of Educational Studies
University of Delaware

Paper presented at the annual meeting of the National Council on Measurement in Education, April 19, 1995.

TM024594

BEST COPY AVAILABLE

Application of SIBTEST in Dealing With Issues of DIF in the Context of Multidimensional Data

Abstract

A modification of SIBTEST to assess DIF for two-dimensional test data is described. That is, there are two intended abilities tapped by test items. A small simulation study is carried out to assess the performance of modified SIBTEST to detect DIF in such two-dimensional data. The simulation study results indicate that the modified SIBTEST has exhibited good adherence to the nominal level of significance for cases where there are two intended abilities with no DIF present, while exhibiting high power for cases where there are two intended abilities with DIF present. The results are discussed with implications for future research.

The topic of Differential item functioning (DIF) has attracted many researchers in recent years. At the item level, an item is said to exhibit DIF if the item has different probabilities of correct response for examinees of similar ability but belonging to different groups (usually based on race, gender, socioeconomic status, etc.), generally referred to as focal (F) and reference (R) groups. A group of items is said to exhibit differential testlet functioning (DTF) if the expected scores on this subtest differs for examinees of similar ability but belonging to different groups. Several methodologies have been developed to assess DIF/DTF and numerous studies have been carried out to study the validity of these methodologies to assess DIF. Notable among these methodologies are: Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), SIBTEST procedure (Shealy & Stout, 1993a, 1993b), logistic regression procedure (Swaminathan & Rogers, 1990), and item response theory (IRT) based procedures (Shepard, Camilli, & Williams, 1985). Among the procedures mentioned here SIBTEST is the only procedure that has the capability to assess both DIF and DTF. Other procedures can assess only DIF.

Most studies conducted thus far have studied DIF/DTF of unidimensional tests. That is, it was presumed that the intent of the test was to assess a unidimensional ability. In order to assess DIF/DTF in unidimensional tests, the general procedure is to divide examinees from the two groups (R and F) into subgroups based on a single score derived from the test items (referred to as the matching subtest). In each of the subgroups, the performance of examinees from R and F groups is compared on the item(s) of interest and aggregated across subgroups to arrive at a statistic that is used to assess DIF/DTF. Few studies have investigated DIF in the context of multidimensionality. That is, in the case of multidimensionality, it is presumed that the intent of the test is to measure more than one dominant ability. Oshima and Miller (1992) have shown that when test data has been produced by two latent abilities, and if the total test score is used as the matching subtest, then IRT based DIF indices were able to detect DIF items effectively provided the contamination of multidimensionality was small (up to five percent). The power of these indices to detect DIF decreased as the degree of multidimensionality increased. Ackerman (1992) also has shown that using the total test score of a multidimensional test as the matching criterion could lead to high type–I error rates. These studies illustrate that using total test score as a matching criterion in cases of multidimensional tests could result in intentional latent multidimensionality manifesting itself as DIF.

In applications, often there are situations when the tests are deliberately constructed multidimensional. That is, items tap more than one dominant dimension. In such situations it is unrealistic to presume that a unidimensional test score will adequately represent a valid matching subtest. For example, in a general science test or an advanced placement test, generally there is more than one dominant ability influencing examinee

responses. In these cases, using total test score as a matching criterion could lead to misleading results. For expository purposes, let us suppose that in a two-dimensional test, half the items tap ability1 (θ_1), and the other half tap ability2 (θ_2). Then, if you look at a subgroup consisting of examinees with 50% correct on the total test, it is evident that one could obtain 50% correct on the test in several ways: by answering all of the items tapping (θ_1), or all of the items tapping (θ_2), or any combination of two sets of items. If we are comparing performance of R and F groups on an item based on such a matching subtest, it could lead to inconsistent and misleading results. Therefore, in order to obtain accurate results, it is important to match examinees so that they are comparable in the intended abilities.

Most existing procedures can only be used for assessing DIF/DTF of unidimensional tests. In order to utilize them for assessing DIF/DTF in the context of multidimensionality, these procedures need to be modified and/or new procedures need to be developed. For example, Mazor, Kanjee, and Clauser (1993), have illustrated the use of logistic regression procedure ,which can be naturally extended to multidimensional tests, to detect DIF in the context of multidimensionality.

The purpose of the present study is to illustrate the application of a modification of SIBTEST to detect DIF/DTF in the context of multidimensionality. I will first describe the SIBTEST procedure for unidimensional tests and then extend it to multidimensional tests.

The SIBTEST Procedure for Unidimensional Tests

It is assumed that two groups of examinees, the reference group (R), and the focal group (F) are administered a given set of items. These items are assumed to be measuring an intended ability denoted by θ, known as the *target ability*. Unintended abilities that influence item responses are denoted by ηs, and are known as *nuisance abilities*. It is also assumed that most items are influenced by one dominant ability, namely the target ability, and that a few items are possibly influenced by more than one dominant ability, namely the target ability and the nuisance ability (or abilities). Items influenced by more than one ability are considered to be potentially exhibiting DIF. For example, in a mathematics reasoning test, some items could be tapping the knowledge or familiarity with baseball, while some other items could be tapping the knowledge of chess in addition to knowledge in mathematics. In this instance, there are two nuisance abilities, knowledge of baseball (η_1), and knowledge of chess (η_2). Knowledge of mathematics is the target ability (θ).

The statistical hypothesis for testing the absence of DIF can be stated as

$$H_0 : \beta_U = 0 \; vs. \; H_1 : \beta_U \neq 0$$

where β_U is the parameter denoting the amount of unidirectional DIF when a single item is considered or unidimensional DTF when more than one item is considered, which is given by

$$\beta_U = \int_\theta B(\theta) f_F(\theta) d\theta,$$

where

$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta)$$

is the difference in the studied response functions for the R and F groups.

The procedure to test for DIF using SIBTEST involves splitting of test items into two subtests: a "matching subtest" and a "studied subtest". The studied subtest contains potentially DIF item(s) while the matching subtest, often but not always, contains the rest of the items. The matching subtest is used to group examinees in the two groups (R and F) into subgroups so that within each subgroup R and F group examinees are comparable on the intended ability.

Let N denote the total number of items. Let items 1 ... n denote matching subtest items, and items n+1 ... N denote studied subtest items. Let U_i denote the response to the item i taking values 0 or 1. For each examinee, let $X = \sum_{i=0}^{n} U_i$ denote the total score on the matching subtest and $Y = \sum_{i=n+1}^{N} U_i$ denote the total score on the studied subtest. Examinees within the reference and the focal groups are grouped into K subgroups (maximum K=n+1) according to their score on the matching subtest. Examinees with the same score on the matching subtest are compared across the reference and the focal groups on their performance on the studied subtest items. The weighted mean difference between the reference and the focal groups on the studied subtest score across K subgroups is then given by

$$\hat{\beta}_U = \sum_{k=0}^{K} \hat{p}_k (\overline{Y}_{Rk} - \overline{Y}_{Fk}) \qquad (1)$$

where $\hat{p}_k$ denotes the proportion of focal group examinees in the subgroup k [1]. Here the $\overline{Y}_{Rk}$ and $\overline{Y}_{Fk}$ denote the adjusted means of the studied subtest (adjusted for mean differences, if there is any, between the reference and the focal groups, using a regression correction as described below), for examinees in the reference and the focal groups

[1] one could also pool examinees from both groups of the kth subgroup.

respectively for subgroup k. $\hat{\beta}_U$ can also be used as a statistic that estimates the amount of unidirectional DIF or DTF. For example a $\hat{\beta}_U$ value of .1 denotes that the average difference between the focal and the reference groups on the studied subtest item(s) is one tenth of a point when controlled for their ability differences. Significant positive values of $\hat{\beta}_U$ denote DIF against the focal group examinees and significant negative values of $\hat{\beta}_U$ denote DIF against the reference group examinees. The test statistic for testing the null hypothesis of no DIF or DTF is given by

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \tag{2}$$

where $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$ given by

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=1}^{K} \hat{p}_k^2 \left(\frac{1}{J_{Rk}} \hat{\sigma}^2 (Y|k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2 (Y|k, F) \right) \right)^{1/2},$$

where J_{Rk} and J_{Fk} denote the number of examinees in the reference and the focal groups respectively in the subgroup k, and $\hat{\sigma}^2 s$ are the respective variances of the studied subtest score.

The null hypothesis is rejected at error rate α if B exceeds the upper $100(1-\alpha)$th percentile point of the standard normal distribution.

The performance of SIBTEST for detecting the unidirectional DIF has been studied extensively through Monte Carlo simulations by various researchers on various factors: the type–I error rate, power, sample sizes, varying degrees of DIF induced, and for varying degrees of ability differences between the groups. (Ackerman, 1993; Narayanan & Swaminathan, 1993, Shealy & Stout, 1993b; Roussos & Stout, 1993). It has been found that the test statistic B has good adherence to the nominal type–I error with high power. The SIBTEST procedure can be applied to sample sizes as small as 100 for each group, and it has exhibited high power for unequal ability distributions between the reference and the focal groups.

The performance of SIBTEST to assess DTF has been studied by Nandakumar (1993) for both simulated and real data sets. In her study both the amplification of DIF due to one nuisance ability and cancellation of DIF, partially or fully, due to two nuisance abilities were studied. Douglas, Li, Roussos, and Stout (1994) have further established the usefulness of SIBTEST to detect differential testlet functioning. In addition, SIBTEST has been expanded to detect crossing DIF (in addition to the uniform DIF), and to deal with polytomous items (Chang, Mazzeo, & Roussos, 1993; Li & Stout,

1993). Finally, a smoothed version of SIBTEST that assesses local DIF especially well has been developed by Douglas, Stout, and DiBello (1994).

Regression Correction to Adjust for Mean Differences on the Target Ability Distribution

Often the target ability distribution of the focal group is lower than that of the reference group. This results in incompatibility in the average scores of the matching subtest for the two groups within subgroup k even when no DIF is present. SIBTEST deals with this problem by estimating the true score of the matching subtest score in the subgroup k for each of the groups (R & F). This amounts to adjusting the observed scores on the matching subtests due to their unreliability of measurement. The corresponding adjusted mean scores on the studied subtest $\overline{Y}^*_{Rk}$ and $\overline{Y}^*_{Fk}$ of Equation 1 are obtained using a Taylor series approach to adjust for F and R group differences in estimated true scores for subgroup k. The details of obtaining the adjusted means are briefly described below (for details see Shealy & Stout, 1993b).

Let X=k denote a matching subtest score. Let $V_g(k)$ denote the regression of the true score on the matching subtest score k on X, and $\hat{V}_g(k)$ its estimate for group g. Also, for subgroup k, it is assumed that $V_g(k)$ is approximately linear in k, leading to

$$V_g(k) = \alpha + \beta k \tag{3}$$

Using the standard regression theory estimates, $V_g(k)$ can be written as

$$V_g(k) = ET + (\frac{\rho_{XT}\sigma_T}{\sigma_X})(k - EX) \tag{4}$$

where T denotes the true score of the target ability as measured by the matching subtest for the subgroup k, and EX denotes the average matching subtest score of examinees in group g (R or F). It is well known from the true-score model (Lord & Novick, 1968, p.61) that

$$\frac{\rho_{XT}\sigma_T}{\sigma_X} = 1 - \frac{\sigma_e^2}{\sigma_x^2},$$

and

$$ET = EX.$$

Therefore, Equation 4 can be written as

$$V_g(k) = EX + (1 - \frac{\sigma_e^2}{\sigma_x^2})(K - EX). \tag{5}$$

$\hat{V}_g(k)$ can be computed by

$$\hat{V}_g(k) = \bar{X}_g + (1 - \frac{\hat{\sigma}^2_{(e|g)}}{\hat{\sigma}^2_{(x|g)}})(k - \bar{X}_g). \tag{6}$$

Here

$$\hat{\sigma}^2_{(x|g)} = \frac{1}{(J_g - 1)} \sum_j^{J_g} (X_{gj} - \bar{X}_g)^2,$$

and

$$\hat{\sigma}^2_{(e|g)} = \sum_{i=1}^{n} \bar{U}_{ig}(1 - \bar{U}_{ig}),$$

where J_g denotes the number of group g examinees taking the test and X_{gj} is the matching subtest score for the jth examinee in group g; $\bar{U}_{ig}$ denotes the proportion correct for group g examinees for the matching subtest item i.

$\hat{V}_g(k)$ in Equation 6 is an estimate of the true score of the target ability for examinees with a score of X=k on the matching subtest from group g (R or F). The corresponding score on the studied subtest can be given, from a Taylor series viewpoint, by

$$\bar{Y}^*_{gk} = \bar{Y}_{gk} + \hat{M}_{gk}(\hat{V}(k) - \hat{V}_g(k)), \tag{7}$$

where $\bar{Y}_{gk}$ is the observed mean on the studied subtest for examinees in group g (R or F) with X=k on the matching subtest and

$$\hat{V}(k) = \frac{1}{2}(\hat{V}_R(k) + \hat{V}_F(k)) \tag{8}$$

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)}, \tag{9}$$

where $\bar{Y}_{g,k+1}$ is the observed mean on the studied subtest for examinees in the subgroup one above the kth subgroup, and $\bar{Y}_{g,k-1}$ is the mean on the studied subtest for examinees in the subgroup one below the kth subgroup. Similar explanation holds for $\hat{V}_g(k-1)$ and $\hat{V}_g(k+1)$.

$\bar{Y}^*_{gk}$ of Equation 7 is an estimate, for subgroup k, of the studied subtest true score for group g presuming an estimated matching subtest score $\hat{V}(k)$ for both groups.

Extension of SIBTEST Procedure for Two-Dimensional Tests

In a two-dimensional situation it is assumed that the test consists of items designed to measure two intended abilities θ_1 and θ_2. Each item could be tapping either one or both of these abilities. Consistent with the multidimensional perspective on which the SIBTEST procedure is based on, potential for DIF arises only if there is one or more unintended abilities $\eta(s)$ influencing one or more items of the test.

Let n_1 denote items tapping the latent ability θ_1, and forming the matching subtest 1 and n_2 denote items tapping the latent ability θ_2 and forming the matching subtest 2. For expository purposes, we assume that the matching subtests' items influence either θ_1 or θ_2, and not both. Let X_1 and X_2 denote test scores on the matching subtest 1 and the matching subtest 2 respectively. Let Y denote scores on the studied subtest. The idea here is that once the examinees are matched from the reference and the focal subgroups on both X_1 and X_2, their scores on the studied subtest item(s) can be compared to assess if DIF/DTF is caused by these items. As in the unidimensional case, the ability distributions of the reference and focal groups on the target abilities θ_1 and θ_2 need to be considered. If there are mean differences on the target ability distributions (which is often found), then these mean differences (impact) could manifest as DIF. That is, they could result in increasing the number of false positives for SIBTEST (or any other) procedure. Therefore, in order to separate impact from the true DIF, the regression correction is performed. The regression correction proposed by Stout is described below. It is a direct extension of the unidimensional case. Also, it is a first attempt, for which refinements will be made.

Examinees from each of the groups (R and F) are divided into subgroups 0 ... n_1 based on their scores on the matching subtest 1, and into subgroups 0 ... n_2 based on their scores on the matching subtest 2. Examinees from the reference and the focal groups are then matched so that within each subgroup both groups of examinees have similar scores on both the subtests (j=1,2). For example the subgroup (k_1, k_2) consists of examinees from the focal and the reference groups who obtained a score of k_1 on the matching subtest 1 and k_2 on the matching subtest 2, assuming that both target ability distributions are the same for the two groups (R and F). If the target ability distributions are not the same for the two groups, then the regression correction equates the groups, so that we have a common score for both groups on each of the target abilities.

Let $V_{gj}(k_j)$ be the expected true score on the matching subtest j (j=1,2) for group g (R of F) in the subgroup with $X_j = k_j$. This is obtained in the manner identical to what was described in the unidimensional case separately for each of the subtests (j=1,2).

$\hat{V}_{g2}(k_j)$s can be computed by

$$\hat{V}_{g1}(k_1) = \bar{X}_{g1} + (1 - \frac{\hat{\sigma}^2_{(e1|g1)}}{\hat{\sigma}^2_{(x_1|g1)}})(k_1 - \bar{X}_{g1}) \tag{10}$$

$$\hat{V}_{g2}(k_2) = \bar{X}_{g2} + (1 - \frac{\hat{\sigma}^2_{(e2|g2)}}{\hat{\sigma}^2_{(x_2|g2)}})(k_2 - \bar{X}_{g2}). \tag{11}$$

Analgous to Equation 8 in the unidimensional case, common matching subtest scores for the reference and the focal groups for the subgroup (k_1, k_2) are defined by

$$\hat{V}_j(k_j) = \frac{\hat{V}_{Rj}(k_j) + \hat{V}_{Fj}(k_j)}{2}, j = 1, 2 \tag{12}$$

The corrected studied subtest score for examinees in the subgroup (k_1, k_2), using a two-dimensional Taylor series approach, is then given by

$$\bar{Y}^{*}_{g(k_1,k_2)} = \bar{Y}_{g(k_1,k_2)} + \hat{M}_{g1(k_1,k_2)}(\hat{V}_1(k_1) - \hat{V}_{g1}(k_1)) + \hat{M}_{g2(k_1,k_2)}(\hat{V}_2(k_2) - \hat{V}_{g2}(k_2)), \tag{13}$$

where $\hat{V}$s are described in Equations 10, 11, and 12, and

$$\hat{M}_{g1(k_1,k_2)} = \frac{\bar{Y}_g(k_1 + 1, k_2) - \bar{Y}_g(k_1 - 1, k_2)}{\hat{V}_{g1}(k_1 + 1) - \hat{V}_{g1}(k_1 - 1)}, g = R, F. \tag{14}$$

$$\hat{M}_{g2(k_1,k_2)} = \frac{\bar{Y}_g(k_1, k_2 + 1) - \bar{Y}_g(k_1, k_2 - 1)}{\hat{V}_{g2}(k_2 + 1) - \hat{V}_{g2}(k_2 - 1)}, g = R, F. \tag{15}$$

Analgous to Equation 9, $\bar{Y}_g$s are the observed means on the studied subtest for examinees in group g in the corresponding subgroups.

Simulation Study

A small scale simulation study was conducted to assess the performance of SIBTEST as extended for two dimensional test data. The matching subtest length was fixed to 40 items, out of which the first 20 items strictly measured θ_1 and the last 20 items strictly measured θ_2. The studied subtest consisted of one item. The examinee sample sizes (J_R and J_F) were each fixed to 3000. Two values of guessing were used, 0 and

.2. Several different values were used for the difference between the means of the target ability distributions ($d_T = \mu_{\theta R} - \mu_{\theta F}$): 0, .25, .3, .4, .5, and 1.0.

Abilities were generated from a bivariate normal distribution with correlation between all pairs of abilities set to 0.5. Two dimensional item parameters for the matching subtest were generated from normal distributions with the following parameters: $\mu_a = 1.22$, $\mu_b = 0$, $\sigma_a^2 = .65$, $\sigma_b^2 = 0.65$. For the type-I error study, item parameters of the studied item had nonzero parameters ($a_{1\theta}$, $a_{2\theta}$, $b_{1\theta}$, $b_{2\theta}$) for either one or both of the abilities. Whereas, for the power study, the studied item had nonzero item parameters corresponding to θ_1, θ_2, and η. Item parameters of the studied subtest item are listed along with results in Tables 1 and 2. For the power study DIF was introduced through the distance (in standard deviation units) between the means on the nuisance ability distributions for the reference and the focal groups (C_β), which is determined by the following equation:

$$C_\beta = E_R(\eta_{\theta 1,\theta 2}) - E_F(\eta_{\theta 1,\theta 2}) = (\mu_{\eta R} - \mu_{\eta F}) + \rho(\mu_{\theta 1 R} - \mu_{\theta 1 F}) + \rho(\mu_{\theta 2 R} - \mu_{\theta 2 F}).$$

Several different values were considered for C_β in simulations: 0.16, 0.23, 0.33, 0.3, 0.4, 0.5, and 1.0.

Item responses were generated using the multidimensional compensatory abilities model due to Reckase and McKinley (1983). The model used to generate responses of the matching subtest items is given by

$$P_i(\theta_1, \theta_2) = c_i + \frac{1 - c_i}{1 + exp(-1.7(a_{i,\theta_1}(\theta_1 - b_{i,\theta_1}) + a_{i,\theta_2}(\theta_2 - b_{i,\theta_2})))}, i = 1, \ldots, n_j \quad (16)$$

and the model used to generate the responses of the studied subtest item is given by

$$P_i(\theta_1, \theta_2, \eta) = c_i + \frac{1 - c_i}{1 + exp(-1.7(a_{i,\theta_1}(\theta_1 - b_{i,\theta_1}) + a_{i,\theta_2}(\theta_2 - b_{i,\theta_2}) + a_{i,\eta}(\eta - b_{i,\eta})))} \quad (17)$$

For each simulated examinee, the probability of correct response for each item was computed using either Equation 12 or Equation 13. If the computed probability was greater than the uniform random variable generated from the interval (0,1), the item was considered answered correctly and a score of 1 was assigned. Otherwise a score of 0 was assigned.

Results

The results of the simulation study are reported in Tables 1 and 2. Table 1 shows results of type-I error study. The first column of Table 1 shows d_Ts for θ_1, θ_2, and η. Columns 2 through 5 show item parameters of the studied item. The last two columns give the estimated value of unidirectional DIF ($\hat{\beta}_U$), and the number of rejections of H_0 out of 100 replications.

As can be seen from the Table 1, rejection rates, out of 100 trials, are around .05 (with the exception of two large values), which is the specified significance level. The average observed type-I error rate is .057. This study has considered mean differences on the target abilities between the groups as large as 1.0, and the presence of guessing. Given these facts, preliminary results look very promising.

The power study is described in Table 2. The first column of Table 2 shows d_Ts. Columns 2 through 8 show item parameters of the studied item. Column 9 shows the potential for DIF induced (C_β), column 10 shows the estimated value of unidirectional DIF ($\hat{\beta}_U$), and the last column gives the number of rejections of H_0 out of 100 replications.

As can be seen from Table 2, generally speaking, the power is good. As the potential for DIF (C_β) increases from .16 to 1.0, the power also increases. This point is further illustrated in rows 11 through 15, where the item parameters are fixed and C_β is varied from .23 to .5. The power goes up from 58% to 100%. The power is much higher for cases where guessing is absent than the corresponding cases where guessing is present.

Discussion

As indicated earlier, the results reported here are preliminary. This is a first attempt made by Stout to extend the SIBTEST procedure to assess DIF for intentionally multidimensional tests. The regression correction as proposed here involves regressing true scores of matching subtests upon observed scores of matching subtests for R and F groups. In this regard, Equations 13, 14, and 15 play an important role. The weights $\hat{M}_{g1}$ and $\hat{M}_{g2}$ of Equation 13 as given in Equations 14 and 15 show that true scores $\hat{V}_{g1}$ and $\hat{V}_{g2}$ of matching subtests 1 and 2 are estimated independently of each other. That is, in the denominator of Equation 14, matching subtest true scores $\hat{V}_{g1}$ at $X_1 = k_1 + 1$ and at $X_1 = k_1 - 1$ are estimated independently of X_2 values. In contrast, the $\hat{Y}$s of Equations 14 are not computed independently of X_2 values. That is, in the numerator

of Equation 14, $\bar{Y}_g(k_1+1, k_2)$ and $\bar{Y}_g(k_1-1, k_2)$ denote means for group g (R or F) on the studied subtest for the subgroups with $X_1 = k_1 + 1$ and $X_1 = k_1 - 1$ on matching subtests 1, but for a fixed score of $X_2 = k_2$ on the matching subtest 2. A similar statement holds for Equation 15. Therefore, the numerators of Equations 14 and 15 are estimated conditionally, while the denominators are estimated unconditionally.

There are alternative ways of estimating the M_gs. For example, the true score of X_1 can be estimated conditional upon $X_2 = k_2$, or the true score of $X_1 = k_1$ at $X_2 = k_2$ can be estimated using examinees "close" to $X_2 = k_2$, and so on. In other words, the numerators of $\hat{M}_g$s can be slightly smoothed. Similarly, the denominators can be slightly restricted, that is, incur less smoothing. These alternative smoothing techniques for regression correction are being pursued currently.

The results presented here are limited in other ways. In the simulation study presented here, items loaded on either of the abilities (θ_1 or θ_2), but not both. In applications, however, one expects to find items influenced by both abilities (θ_1 and θ_2) simultaneously. When items load on more than one ability, appropriate matching subtest scores and the associated regression correction becomes a more challenging problem. Also, in the simulation study conducted here, only one sample size, one test size, and one nuisance ability were used. Future studies will vary each of these as a factor. Also it will be interesting to study amplification and cancellation of DIF due to more than one nuisance ability. Further, the issue of how one chooses the matching subtest is of vital importance in a multidimensional DIF situation, and requires further research.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Ackerman, T. (1993, April). *A didactic example of the influence of conditioning on the complete latent ability space when performing DIF analyses.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). *Extension of Shealy-Stout's DIF procedure to polytomous scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Douglas, J., Li, H., Roussos, L., & Stout, W. (1994). *Testlet bias: Identifying suspect testlets and assessing testlet bias.* Paper submitted for publication.

Douglas, J., Stout, W., & DiBello, L. (1994). *Kernal smoothing version of SIBTEST with applications to local DIF inference and function estimation.* Under review.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Li, H., & Stout, W. (1993, June). *A new procedure for detection of crossing DIF/bias.* . Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA:Addison-Wesley.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Narayanan, P., & Swaminathan, H. (1993, April). *Performance of the Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning.* Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Roussos, L., & Stout, W.(1993, June). *Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance.* Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Shealy, R., & Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shealy, R., & Stout, W. (1993b). An item response theory model for test bias. In P. W. Holland and H.Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Table 1
Results of Type-I Error Study
$J_F = 3000$, $J_R = 3000$

d_T θ_1,θ_2,η	$a_{1\theta}$	$b_{1\theta}$	$a_{2\theta}$	$b_{2\theta}$	c	$\hat{\beta}_U$	#rej
0, 0, 0	0.4	-1.5			0	.001	6
0, 0, 0			1.0	0.5	.2	.002	8
0, 0, 0	1.0	-0.5	1.0	-0.5	.2	.000	4
0, 0, 0	0.4	0.5			0	-.001	4
0, 0, 0			1.0	-1.5	0	-.001	4
.5, .5, 0	0.4	0.5	0.4	0.5	.2	-.003	5
.5, .5, 0	1.0	0.5			.0	-.003	8
.5, .5, 0			2.5	-1.5	.2	.000	4
.5, .5, 0	0.4	-1.5	0.4	-1.5	0	-.002	10
.5, .5, 0	1.0	0.5			0	-.002	3
1.0, 1.0, 0			0.4	0.5	0	-.003	3
1.0, 1.0, 0	0.8	-0.75	0.8	-0.75	.2	-.008	9
1.0, 1.0, 0	2.5	-1.5			0	.001	6
1.0, 1.0, 0			0.4	0.5	.2	-.005	6

Table 2
Results of Power Study
J_F = 3000, J_R = 3000

d_T θ_1,θ_2,η	$a_{1\theta}$	$b_{1\theta}$	$a_{2\theta}$	$b_{2\theta}$	a_η	b_η	c	C_β	$\hat{\beta}_U$	#rej
.5, .5, .5	0.5	0.0	0.5	0.0	1.5	0.0	0	0.16	.041	95
.5, .5, .5	0.5	0.0	0.5	0.0	1.5	0.0	.2	0.16	.033	80
.5, .5, .5	0.5	0.5	1.0	-1.0	1.0	0.0	0	0.16	.027	55
.5, .5, .5	0.5	0.5	1.0	-1.0	1.0	0.0	.2	0.16	.017	27
.25, .25, .5	0.5	0.5	1.0	-1.0	1.0	0.0	0	.33	.059	100
.25, .25, .5	0.5	0.5	1.0	-1.0	1.0	0.0	.2	.33	.047	100
.25, .25, .5	0.5	0.0	1.0	0.0	0.5	0.0	0	.33	.033	88
.25, .25, .5	0.5	0.0	1.0	0.0	0.5	0.0	.2	.33	.027	60
.25, .25, .5	1.0	0.0	1.0	0.0	0.5	0.0	0	.33	.032	88
.25, .25, .5	1.0	0.0	1.0	0.0	0.5	0.0	.2	.33	.019	33
.4, .4, .5	2.5	-1.5	0.0	0.0	1.0	0.0	0	.23	.014	58
.3, .3, .5	2.5	-1.5	0.0	0.0	1.0	0.0	0	.3	.017	80
.25, .25, .5	2.5	-1.5	0.0	0.0	1.0	0.0	0	.33	.018	79
.3, .3, .6	2.5	-1.5,	0.0	0.0	1.0	0.0	0	.4	.024	97
.3, .3, .7	2.5	-1.5	0.0	0.0	1.0	0.0	0	.5	.030	100
0, 0, .5	1.0	0.0	1.0	0.0	1.0	0.5	0	0.50	.079	100
0, 0, .5	1.0	0.0	1.0	0.0	1.0	0.5	.2	0.50	.065	100
0, 0, 1.0	1.0	0.0	1.0	0.0	1.0	0.5	0	1.0	.158	100
0, 0, 1.0	1.0	0.0	1.0	0.0	1.0	0.5	.2	1.0	.129	100